TRIC: an automated alignment strategy for reproducible protein quantification in targeted proteomics

Röst, H. L., Liu, Y., D'Agostino, G., Zanella, M., Navarro, P., Rosenberger, G., Collins, B. C., Gillet, L., Testa, G., Malmström, L., & Aebersold, R. (2016). TRIC: an automated alignment strategy for reproducible protein quantification in targeted proteomics. *Nature Methods*, *13*(9), 777-783. https://doi.org/10.1038/nmeth.3954

Published in:
Nature Methods

Document Version:
Peer reviewed version

Reproducible protein quantification with TRIC: An automated alignment strategy for comprehensive data matrices in targeted proteomics

Hannes L. Röst,[1,2] Yansheng Liu,[1] Giuseppe D'Agostino,[3] Matteo Zanella,[3] Pedro Navarro,[1,4] George Rosenberger,[1,5] Ben C. Collins,[1] Ludovic Gillet,[1] Giuseppe Testa,[6,3] Lars Malmström,[1] and Ruedi Aebersold[1,7,*]

[1]*Department of Biology, Institute of Molecular Systems Biology, ETH Zurich, Zurich, Switzerland*

[2]*Department of Genetics, Stanford University, Stanford, CA, USA*

[3]*Department of Experimental Oncology, European Institute of Oncology (Istituto di Ricovero e Cura a Carattere Scientifico, IRCCS), Milan, Italy*

[4]*Institute for Immunology, University Medical Center of the Johannes Gutenberg University Mainz, Mainz, Germany*

[5]*Ph.D. Program in Systems Biology, University of Zurich and ETH Zurich, Zurich, Switzerland*

[6]*Department of Oncology and Hemato-Oncology, University of Milan, Italy*

[7]*Faculty of Science, University of Zurich, Zurich, Switzerland*

(Dated: May 18, 2016)

Abstract

Large scale, quantitative proteomics studies have become essential for the analysis of clinical cohorts, large perturbation experiments and systems biology studies. While next-generation mass spectrometric techniques such as SWATH-MS have substantially increased throughput and reproducibility, ensuring consistent quantification of thousands of peptide analytes across multiple LC-MS/MS runs remains a challenging and laborious manual process. To produce highly consistent and quantitatively accurate proteomics data matrices in an automated fashion, we have developed the TRIC software which utilizes fragment ion data to perform cross-run alignment, consistent peak-picking and quantification for high throughput targeted proteomics. TRIC uses a graph-based alignment strategy based on non-linear retention time correction to integrate peak elution information from all LC-MS/MS runs acquired in a study. When compared to state-of-the-art SWATH-MS data analysis, the algorithm was able to reduce the identification error by more than 3-fold at constant recall, while correcting for highly non-linear chromatographic effects. On a pulsed-SILAC experiment performed on human induced pluripotent stem (iPS) cells, TRIC was able to automatically align and quantify thousands of light and heavy isotopic peak groups and substantially increased the quantitative completeness and biological information in the data, providing insights into protein dynamics of iPS cells. Overall, this study demonstrates the importance of consistent quantification in highly challenging experimental setups, and proposes an algorithm to automate this task, constituting the last missing piece in a pipeline for automated analysis of massively parallel targeted proteomics datasets.

*Corresponding author: aebersold@imsb.biol.ethz.ch; Phone: +41 44 633 31 70; Fax: +41 44 633 10 51

Introduction

Molecular biology is increasingly becoming a data-driven science, which enables researches in biology and medicine to investigate large numbers of biological systems on a genome-wide scale. Underlying this transition is the ability to generate robust, comprehensive and fully quantitative "data matrices" capturing measurements across many samples (first dimension) in a genome-wide fashion (second dimension). In nucleic acid sequencing-based fields, this transition has advanced enough to allow for large-scale inference from thousands of samples in a reproducible and comparable manner [1–5].

In contrast, in the field of proteomics the transition to high-throughput measurements across large numbers of samples has proven to be challenging (**Supplementary Note 1** and Röst et al. [6]). While discovery-oriented techniques, such as data-dependent acquisition (DDA) [7–10], have recently allowed the identification of a large part of the human proteome [11, 12], it has become apparent that these methods suffer from poor reproducibility in large scale experiments. Particularly when applied in high throughput to complex protein mixtures, e.g. whole proteomes, the resulting data matrices contain many missing values. To improve reproducibility, alternative approaches based on targeted proteomics were developed, which provide high consistency and quantitative accuracy across many experimental conditions due to their deterministic acquisition strategy. Specifically, selected reaction monitoring (SRM) proved to be invaluable for large-scale measurements geared towards systems biology [13] or biomarker discovery [14–17]. However, while SRM-based targeted proteomics produces highly consistent data matrices, it is limited by low throughput, resulting in output matrices with typically only few tens of quantified proteins per study (**Supplementary Fig. 1**).

Recently, we developed SWATH-MS based on the principle of targeted analysis of data-independent acquisition (DIA) data as a method for massively parallel targeted proteomics [18]. Our targeted analysis of DIA data based on OpenSWATH was able to increase the throughput of targeted proteomics by several orders of magnitude compared to SRM-based approaches, and is, in principle, able to to generate proteome-wide data matrices [6, 19]. However, obtaining consistent and accurate matrices from targeted proteomics data is challenging as most current software was developed for low-throughput SRM data and focused on manual analysis and visualization of the data [20–26]. Even fully automated

software solutions for peak picking and error rate estimation [19, 27, 28] generally only operate on a single MS run at a time and are unable to efficiently integrate experimental information from multiple targeted MS runs. However, a single MS run may not contain sufficient information to confidently select the correct peptide elution time point among multiple detected peak groups of similar quality in a given chromatogram (**Fig. 1 a**). Analyzing single MS runs in isolation, therefore, cannot ensure consistent peak picking across a whole experiment (**Supplementary Note 2**).

Here we describe the TRIC (TRansfer of Identification Confidence) algorithm, an automated method which integrates all information from a targeted proteomics experiment to accurately and consistently determine the correct elution peak in each MS run. The software is designed to exploit the particular structure of chromatographic fragment ion-based peak data found in targeted proteomics, providing quantification and identification in the same algorithm. The TRIC algorithm uses a reference-free alignment approach based on individual, pairwise non-linear retention time (RT) de-warping, which allows it to scale to hundreds of targeted proteomics LC-MS/MS runs. Together with the OpenSWATH framework [19], TRIC allows fully automated analysis of next-generation targeted proteomics datasets with high throughput. The software is vendor-independent and provided as an open-source package (Modified BSD Licence) at `https://pypi.python.org/pypi/msproteomicstools`.

Results

Design and Structure of the TRIC Algorithm

The TRIC algorithm was developed to perform integrative analyses of a large number of targeted MS injections with a focus on robustness, scalability and performance (**Supplementary Note 3**). Inherently, the algorithm was designed to correct for non-linear chromatographic distortion between the runs, minimize the effect of outlier runs, and be scalable to hundreds of MS runs (**Supplementary Note 4**). This is achieved through a guidance tree (**Supplementary Fig. 2**) learned from the data which removes the need for a reference run in alignment and by using a locally adaptive retention time tolerance for each pairwise alignment [29, 30]. The algorithm does not rely on MS1 measurements but works directly on chromatographic peak identifications from primary identification tools, such as

OpenSWATH [19] or PeakView, which usually identify multiple potential peakgroups in an extracted fragment ion chromatogram. Using the fragment ion data, the algorithm is able to boost identification confidence for peaks consistently detected across multiple runs and can provide improved peak boundaries for analyte quantification. Starting from a typical targeted proteomics dataset with a number of individually analyzed and scored runs containing multiple potential peak group candidates for identification (**Fig. 1 a**), the algorithm performs the following steps:

(a) *Alignment*: Using a set of high-confidence endogenous peptides, an estimate of the pairwise chromatographic distance between all MS runs is obtained and used to generate a guidance tree (**Fig. 1 b**). Then, for each edge in the tree, a pairwise non-linear transformation between the RT domains of the two MS runs (nodes) is computed (step I). Note that TRIC does not rely on spike-in peptides and that the guidance tree removes the need for a reference run and thus prevents chromatographically dissimilar runs to be aligned (**Supplementary Figs. 3** and **4**).

(b) *Confidence transfer*: Traversal of the global guidance tree starts for each measured peptide (targeted proteomics assay) from a suitable "seed" (**Fig. 1 b**; step II). During traversal, each node (LC-MS/MS run) of the tree is visited sequentially (step III, iterations 2 and 3) and a confident identification is mapped from one node (run n) to an adjacent node (run m), where our choice of guidance tree ensures that the mapping only occurs between chromatographically similar runs (**Supplementary Note 3**). During confidence transfer, the identification confidence of all peakgroups in run m within the adaptive retention time window (area indicated in gray) is increased; if the confidence score of the best peakgroup passes the user-defined threshold, it gets added to the result. The approach automatically adopts the retention time window for different parts of the tree, depending on the quality of the pair-wise alignment, thus increasing robustness and decreasing the influence of outlier runs (**Supplementary Figs. 5** and **6**).

(c) *Re-quantification*: In the last (optional) step, nodes in the guidance tree where no peakgroup passed the confidence filter can be re-visited for re-quantification. In these cases, the software can infer the peak boundaries from the closest neighboring node and quantify the fragment ion signal within those boundaries. These imputed values, however, are *not* substitutes for quantification events, but merely serve as upper bounds of the analyte signal for the node in question (orange circles, step IV).

In order to control the false discovery rate (FDR), our software also offers the option to perform an error-rate correction based on known false signals ("decoy signals") at the assay level [31], preventing the accumulation of false positive identifications when analyzing multiple runs. This step can remove individual rows from the data matrix if they do not pass the filter criteria (see Methods). This is achieved by requiring a more stringent quality threshold for the "seed" identification, similar to the way the Mayu software [32] operates.

Technical Validation

To validate the alignment and FDR control approach implemented in TRIC, we created a manually validated data set of 7,232 chromatograms which were extracted from the *Streptococcus pyogenes* dataset of Röst et al. [19] (**Supplementary Note 5**). First, 452 peptides were randomly selected from the data, giving rise to 7,232 chromatograms that were loaded into the Skyline software where the correct elution peak, if present, was identified by visual inspection (**Supplementary Table 1**). In parallel, we analyzed the same data with TRIC and compared its performance to the current state-of-the-art, applying a fixed q-value cutoff in each run individually ("naïve approach"). We found that the TRIC algorithm decreased the error rate substantially, while maintaining high recall for peakgroups across all runs (see FDR-Recall behavior in **Fig. 2 a** and **Supplementary Figs. 7–12**). At the same FDR cutoff, TRIC was able to reduce the error rate by more than 3-fold from 1.8 % to 0.5 % (better than the expected 1 %, **Fig. 2 a,b**). Running TRIC with non-linear RT alignment substantially outperformed linear alignment strategies as well as the naïve approach (**Fig. 2 b**).

Next, we evaluated the accuracy of the fragment ion peak boundaries reported by TRIC in the "re-quantification" step. From a dataset of eight *S. pyogenes* runs with large chromatographic differences, we removed 506 high-confidence peakgroups to test whether TRIC could recover them. We observed clear non-linearities in the retention times, which TRIC was able to correct satisfactorily (**Figure 2 c** top *versus* bottom panel). After correction, over 96.6 % of all data fell within $\pm 30\,s$ (less than 1 % of the chromatographic gradient) around the true retention time, compared to 82 % and 47 % for linear or no alignment (**Fig. 2 d** and **Supplementary Fig. 13**). Similarly, we found that 80 % of all reported intensity values deviated less than 25 % from the true intensity value (**Supplementary Fig. 13**). We con-

cluded that the alignment and confidence transfer procedure performed by TRIC improves accuracy, reduces the error rate and correctly accounts for large non-linear chromatographic effects.

Application to microbial virulence

We then applied the TRIC algorithm to the full 12 SWATH-MS runs described in Röst et al. [19], comparing cultures of strain SF370 grown in 0 % and 10 % human plasma to study proteomic changes that occur upon vascular invasion of *S. pyogenes*. TRIC substantially lowered the overall number of missing values as well as the number of incomplete rows generated with each newly added run in the label-free, quantitative proteomics data matrix (**Fig. 3 a**). The guidance tree created by the alignment mostly reflected the biological condition (case *versus* control) and not the acquisition order, while chromatographically dissimilar runs were correctly placed at the periphery of the tree, decreasing their influence on the alignment process (**Fig. 3 b**).

The final assay-level data matrix (**Supplementary Fig. 14**) was to 87 % populated with quantified peakgroups, compared to 69 % using a "naïve approach" with a fixed q-value cutoff of 0.0015 (**Fig. 3 a** and **Supplementary Note 6**). Using the aligned data matrices, we identified 130 *S. pyogenes* proteins that significantly change upon exposure to human plasma (adjusted $p < 0.01$ and effect size larger than 1.5), up by 37 % from 95 proteins without alignment (**Supplementary Tables 2 and 3**). The number of assays quantified in all 12 runs increased by 39 % from 4971 to 6914 (**Fig. 3 c,d**) and substantially fewer peptides were identified in a single run only (down by over 15-fold, compare **Fig. 3 c,d**); TRIC thus added additional quantitation events to singleton peptides (**Supplementary Fig. 15**).

We then investigated whether our algorithm was able to improve the identification consistency and error rate control across the whole experiment. We found that TRIC increased identification consistency across all runs, and the cumulative number of peptide identifications shows early saturation (**Fig. 3 e, f**; see **Supplementary Figs. 16–18** for data on assay level and protein level). This is consistent with a complete mapping of the expressed peptides, while a continuing increase in cumulative identifications, as seen without alignment, would be consistent with an accumulation of false positive identifications.

Protein turnover analysis using TRIC

Protein degradation rates vary substantially across the proteome, and together with transcription rates are the main determinant of the amount of protein in a cell [35, 36]. Protein degradation plays an important role in many cellular processes including cell-cycle, DNA repair, growth and differentiation, and it has been linked to several diseases [37]. To study protein degradation using targeted proteomics, we performed a pulsed-SILAC experiment on induced pluripotent stem cells (iPS) obtained from a healthy human donor, allowing us to measure *in vitro* protein turnover rates in a personalized fashion: After growing the iPS cells in biological duplicates, we replaced the light medium with heavy labeled medium at timepoint zero and samples harvested after 1.5, 4.5 and 13.5 hours were analyzed on an AB Sciex 5600 plus TripleTOF system in SWATH mode (**Fig. 4 b**). Using a matching spectral library, OpenSWATH quantified 5,484 heavy-light pairs mapping to 1,427 proteins (achieving 87 % library coverage for the light precursors).

We reasoned that this dataset will allow us to evaluate our algorithm on a highly heterogeneous time-course dataset with interesting biological applications. First, this setup allowed us to directly test whether TRIC overannotated the resulting data-matrices, i.e. by falsely aligning heavy species at timepoint zero (before heavy amino acids were added). Furthermore, it provided a straight-forward metric to assess the quality of the data by checking the elution time difference error between corresponding heavy and light peptides (the two channels were treated completely independently for the purpose of this analysis). Third, as heavy lysine and arginine get incorporated into the proteome at time point zero, the algorithm will have to accurately quantify very low abundant heavy peptide species–as expected in the early time points–which is very challenging.

Applying TRIC increased the number of quantified SILAC pairs by 62 % and 40 % in the time points 1.5 h and 4.5 h, respectively, while only adding few false positive heavy identifications at time point zero (**Fig. 4 d** and **Supplementary Note 7**). Similarly, the number of quantified proteins detected in fewer than three samples decreased by a factor of 1.9, while the number of quantified proteins identified in five or more samples increased by 59 %. The additional quantification events reported by TRIC increased the error in heavy/light elution only slightly (**Fig. 4 c**). When matched within their respective intensity range, the error distributions are very similar (**Fig. 4 a**, top distribution). Thus, by applying

TRIC to a data structure typically encountered in time-course experiments, the number of quantification events increased by up to 60 % without any significant impact on accuracy.

We then used the SILAC ratios to compute the relative isotopic abundance (RIA) for each peptide over time and fitted an exponential decay model as described by Pratt et al. [38] (see Methods). After filtering and correction for dilution, we obtained the median k_{loss}, the rate of loss of light isotope over time, for 1075 proteins (**Fig. 5 a**). The computed protein-level turnover rates ranged from less than 10 h to several hundred hours with a median protein turnover time of 39.4 h (**Fig. 5 b** and **Supplementary Table 4 and 5**). A gene ontology (GO) enrichment analysis on the proteins with the highest and lowest turnover rates using GORILLA [39] identified 20 significantly (q-value < 0.05) enriched terms (**Supplementary Table 6**); with an enrichment of 3.44-fold, the "cell adhesion" GO term was significantly ($p < 10^{-7}$) enriched in the set of proteins with high turnover (**Fig. 5 c** and **Supplementary Fig. 19**). The enrichment in cell adhesion molecules is both consistent with the hypothesis of a generally faster turnover of these molecules in human cells as well as with the critical involvement of this class of molecules in the regulation of pluripotency (**Supplementary Note 7**). Thus, TRIC enables the accurate identification of protein turnover rates in human iPS cells in a highly challenging time-series experiment.

Discussion

The availability of accurate, consistent and complete data matrices is crucial for systems biology investigations in the field of proteomics. They are the basic currency of data-driven experiments and their accuracy largely determines the success of downstream analyses [6]. The TRIC algorithm described here is capable of creating consistent targeted proteomics data matrices by performing retention time alignment of fragment ion chromatograms and subsequent identification and quantification. The algorithm is specifically designed for targeted proteomics data and works directly with chromatographic peaks identified by upstream tools on MS2 level [19, 27, 33]. By relying on fragment ion-based identification in all runs, TRIC omits the error-prone step of mapping unidentified MS1 features across runs commonly performed in MS1-based alignment software (**Supplementary Note 2**) [29, 30, 41–44]. Instead, the TRIC algorithm employs a "confidence transfer" step where identification confidence (and not the identification itself) is transferred across runs.

The chosen alignment strategy using a globally optimal guidance tree results in minimal alignment error since every alignment step is local and performed between two highly similar runs (as opposed to aligning all runs against a more distant reference run). This makes TRIC scalable to a large number of samples, tolerant to outlier runs and applicable to heterogeneous experimental conditions; which we demonstrate in **Supplementary Note 4** using a dataset with over two hundred blood plasma samples [40]. On this dataset, our reference-free strategy has better precision-recall characteristics than a reference-based approach and the data indicate direct benefits of the adaptive retention time windows (**Supplementary Figs. 3–6**).

Using a validation dataset of 7,232 manually curated ion chromatograms, we find that our algorithm can reach high recall rates while reducing the error rate by a factor of three or more compared to state-of-the-art. In addition, we also compared the TRIC algorithm against DIA-Umpire, an orthogonal algorithm which uses untargeted analysis of DIA to generate pseudo-spectra and then performs RT alignment (**Supplementary Note 5**) [45]. On the manually curated dataset, TRIC achieves higher recall (85 % *versus* 59 %) at lower error rate (0.3 % *versus* 3.8 %) than DIA-Umpire, thus highlighting the benefits of using targeted fragment ion information for identification and alignment (**Supplementary Figs. 11** and **12**). Furthermore, when applying TRIC to large-scale microbial and human targeted proteomics datasets, the number of quantified values consistently increases by 30-60 %, which directly leads to an improvement of statistical power and biological information (**Supplementary Notes 4, 6** and **7**). When applying TRIC to small and large scale data sets (hundreds of injections), we observe consistent performance and scalability to large sample numbers, thus opening the possibility of SWATH-MS to be applied in multi-center studies and, through TRIC, achieve consistent and reproducible research data across labs.

Methods

Experimental procedures

Cell culture The iPSC line used in this study (CTL1R-1) was derived and initially cultured as described in Adamo et al. [46]. Cells were growing in single-cell condition in mTeSR-1 and then adapted for 2 passages in a custom medium, composed as follows: DMEM, Knockout Serum Replacement 15 % (Sigma), Pen-Strep 1 %, Non-essential aminoacids 1 %, Glutamine 1 %, Probumin 0.5 % (Millipore), beta-meracptoethnaol 0.1 mM, L-Proline 500 mg/l (Sigma), FGF2 10 ng/ml (Peprotech). The medium was conditioned for 24 hours on a mouse embryonic fibroblast layer inactivated with mitomycin-C and filtered before use. In the SILAC version of the medium a custom DMEM (Lonza) without arginine and lysine was complemented with 84 mg/l $^{13}C_6$ $^{15}N_4$ Arg10 (Sigma) and 146 mg/l $^{13}C_6$ $^{15}N_2$ Lys8 (Sigma). Cells were scraped and washed in cold PBS upon reaching 70 % confluence approximately for protein harvest.

Cells were counted using a Bürker chamber with Trypan blue counting 5 fields and averaging. Each count was done in duplicate.

MTS assay 20 μl of CellTiter 96 AQueous One Solution reagent (Promega) was added into each well of a 96 multiwell plate containing 5×10^3 cells in 100 μl of mTeSR. Plates were incubated for 1 hour at 37 °C and 490 nm absorbance was recorded using Glomax Multi Detection System (Promega).

Protein extraction and in-solution digestion

The iPSC cell pellets were lysed on ice by using a lysis buffer containing 8 M urea (Euro-Bio), 40 mM Tris-base (Sigma-Aldrich), 10 mM DTT (AppliChem) and complete protease inhibitor cocktail (Roche). The resulted mixture were sonicated in 4 °C for 5 mins using a VialTweeter device (Hielscher-Ultrasound Technology) and centrifuged at 21 130 g and 4 °C for 1 hr to remove the insoluble material. The supernatant protein mixtures were transferred and the protein amount was determined with a Bradford assay (Bio-Rad, Hercules, CA, USA). The protein mixtures was reduced by 5 mM tris(carboxyethyl)phosphine (Sigma-Aldrich) and alkylated by 30 mM iodoacetamide (Sigma-Aldrich). Then 5 volumes of precooled precipitation solution containing 50 % acetone, 50 % ethanol, and 0.1 % acetic

acid was added to the protein mixture and kept at 20 °C overnight. The mixture was centrifuged at 20,400 g for 40 min. The pellets were washed with 100 % acetone and 70 % ethanol with centrifugation at 20,400 g for 40 min. The samples were then resolved by 100 mM NH_4HCO_3 and were digested with sequencing-grade porcine trypsin (Promega) at a protease/protein ratio of 1:40 overnight at 37 °C [47]. Digests were purified with Vydac C18 Silica MicroSpin columns (The Nest Group Inc.). Peptide amount was determined by using Nanodrop ND-1000 (Thermo Scientific) and about 0.7 μg peptide mixtures were analyzed in each LC-MS run. An aliquot of retention time calibration peptides from iRT-Kit (Biognosys) was spiked into each sample before all LC-MS analysis at a ratio of 1:20 (v/v) for linear RT correction in OpenSWATH [48].

Shotgun measurement

The peptides digested from two individual iPSC cells in time zero (cells in light medium) were both measured on an AB SCIEX 5600 plus TripleTOF mass spectrometer operated in DDA mode. The mass spectrometer was interfaced with an Eksigent NanoLC Ultra 2D Plus HPLC system as previously described [49, 50]. Peptides were directly injected onto a 20-cm PicoFrit emitter (New Objective, self-packed to 20 cm with Magic C18 AQ 3-μm 200-Å material), and then separated using a 120-min linear gradient of 2 % buffer B to 35 % buffer B (buffer A 0.1 % (v/v) formic acid, 2 % (v/v) acetonitrile, buffer B 0.1 % (v/v) formic acid, 90 % (v/v) acetonitrile) at a flow rate of 300 nL/min. MS1 spectra were collected in the range 360–1,460 m/z. The 20 most intense precursors with charge state 2–5 which exceeded 250 counts per second were selected for fragmentation, and MS2 spectra were collected in the range 50–2,000 m/z for 100 ms. The precursor ions were dynamically excluded from reselection for 20 s.

SWATH-MS measurement

The same LC-MS/MS systems used for shotgun measurements above was also used for SWATH analysis [49, 50]. Specifically, in the present SWATH-MS mode, the AB SCIEX 5600 plus TripleTOF instrument was specifically tuned to optimize the quadrupole settings for the selection of 64 variable wide precursor ion selection windows. The 64-variable window

schema was optimized based on a normal human cell lysate sample, covering the precursor mass range of 400–1,200 m/z. The effective isolation windows can be considered as being 399.5 to 408.2, 407.2 to 415.8, 414.8 to 422.7, 421.7 to 429.7, 428.7 to 437.3, 436.3 to 444.8, 443.8 to 451.7, 450.7 to 458.7, 457.7 to 466.7, 465.7 to 473.4, 472.4 to 478.3, 477.3 to 485.4, 484.4 to 491.2, 490.2 to 497.7, 496.7 to 504.3, 503.3 to 511.2, 510.2 to 518.2, 517.2 to 525.3, 524.3 to 533.3, 532.3 to 540.3, 539.3 to 546.8, 545.8 to 554.5, 553.5 to 561.8, 560.8 to 568.3, 567.3 to 575.7, 574.7 to 582.3, 581.3 to 588.8, 587.8 to 595.8, 594.8 to 601.8, 600.8 to 608.9, 607.9 to 616.9, 615.9 to 624.8, 623.8 to 632.2, 631.2 to 640.8, 639.8 to 647.9, 646.9 to 654.8, 653.8 to 661.5, 660.5 to 670.3, 669.3 to 678.8, 677.8 to 687.8, 686.8 to 696.9, 695.9 to 706.9, 705.9 to 715.9, 714.9 to 726.2, 725.2 to 737.4, 736.4 to 746.6, 745.6 to 757.5, 756.5 to 767.9, 766.9 to 779.5, 778.5 to 792.9, 791.9 to 807, 806 to 820, 819 to 834.2, 833.2 to 849.4, 848.4 to 866, 865 to 884.4, 883.4 to 899.9, 898.9 to 919, 918 to 942.1, 941.1 to 971.6, 970.6 to 1006, 1005 to 1053, 1052 to 1110.6, 1109.6 to 1200.5 (containing 1 m/z for the window overlap). SWATH MS2 spectra were collected from 50 to 2,000 m/z. The collision energy (CE) was optimized for each window according to the calculation for a charge 2+ ion centered upon the window with a spread of 15 eV. An accumulation time (dwell time) of 50 ms was used for all fragment ion scans in high-sensitivity mode and for each SWATH-MS cycle a survey scan in high-resolution mode was also acquired for 250 ms, resulting in a duty cycle of *ca.* 3.45 s.

Peptide identification for shotgun proteomics

Profile-mode wiff files from shotgun data acquisition were centroided and converted to mzML format using the Sciex Data Converter v.1.3 and converted to mzXML format using msconvert v.3.0.4238 from the proteowizard package [51]. The MS2 spectra were queried against the reviewed canonical SwissProt complete proteome database for human (`ex_sp_9606`, August 2014) appended with reversed sequence decoys [31]. Two types of search engines, X!Tandem [52] and OMSSA [53], were used through iPortal interface for proteomic workflows [54]. The search parameters were: static modifications of 57.02146 Da for cysteines, variable modifications of 15.99491 Da for methionine oxidations. The parent mass tolerance was set to be 50 p.p.m and mono-isotopic fragment mass tolerance was 0.1 Da (which was further filtered to be <0.05 Da for building the spectral library); Fully-

tryptic peptides and peptides with up to two missed cleavages were allowed. The identified peptides were processed and analyzed through Trans-Proteomic Pipeline 4.5.2 (TPP) [55] and were validated using PeptideProphet [56] (PeptideProphet parameter is `-p0 -dDECOY_ -OAPdlIw`). All the peptides were filtered at a false discovery rate (FDR) of 1 % (iProphet probability > 0.8790) [57].

Spectral library generation

The raw spectral libraries were generated as described in Schubert et al. [58] from all valid peptide spectrum matches for the shotgun measurement of the light peptides, and then refined into non redundant consensus libraries [50] using SpectraST [59]. For each peptide, the retention time was mapped into the iRT space [48] with reference to a linear calibration constructed for each shotgun run, as previously described [50]. The light MS assays constructed from the 6 most intense y ions (all b ions and other ions removed) with Q1 range from 400 to 1200 m/z excluding the precursor SWATH window were used for targeted data analysis of SWATH maps. This library was subsequently used to generate corresponding assays for all heavy peptides by shifting the y ion transitions in mass corresponding to the number of lysine and arginine contained in the partial sequence. Decoy assay were appended to the final library as described previously [19].

SWATH-MS data analysis and q-value estimation

SWATH-MS raw data was converted from wiff files to mzXML using msconvert included in the open-source proteowizard package [51]. The OpenSWATH analysis workflow was essentially executed as described in Röst et al. [19] but the improved single executable `OpenSwathWorkflow` was used instead of the multi-step workflow to perform peak-picking and feature detection on all SWATH-MS runs (see `http://www.openswath.org/`). The experimental data and the assay libraries for the *S. pyogenes* samples were obtained from previously published analyses (Röst et al. [19]) while the pulsed-SILAC data was analyzed with a sample-specific library as described above. Both libraries contained target assays as well as decoy assays and we used pyprophet [33], an open-source re-implementation of the mProphet algorithm [27], to perform q-value estimation on individual runs after feature

detection. The algorithm uses semi-supervised machine learning techniques to optimally separate true target assays and known false "decoy" signals and estimate q-values based on their distributions [27]. The reported features and their associated q-values (used as scores here) comprise the input for the TRIC algorithm.

Decoy based FDR control

In a separate step from the actual alignment, our approach allows for the adjustment of the false discovery rate based on a decoy model. This approach can be applied before or after alignment (for convenience, it has been integrated into the TRIC executable). The algorithm applies a q-value threshold based on the score of the best peakgroup per row (since quantification events are likely not independent, this is the most conservative approach). This approach attempts to control the number of decoy *rows* in the output matrix by using the estimate of r, the ratio of false positives to decoys as computed by the Storey-Tibshirani method implemented in mProphet and pyProphet [27, 33]. First, this ratio r is used to estimate the number of false positive rows n_{fp} in the data matrix based on the number of decoy rows n_{decoy} in the matrix: $n_{fp} = r \cdot n_{decoy}$. The q-value threshold is lowered until the desired number of false positives rows (as estimated with the above formula) is reached, for example until $\frac{n_{fp}}{n_{tot}}$ reaches 0.01 (where n_{tot} represents the total number of target rows). Thus, a more stringent score-cutoff is chosen that limits the number of decoy rows in the final data matrix to the user-defined FDR value. In the examples described here, the q-value cutoffs used to achieve a 1 % FDR were 0.0015 for the *S. pyogenes* study, 0.0022 for the pulsed-SILAC study and 2.1×10^{-5} for the blood plasma dataset in the supplementary (generally, the more samples are analyzed, the lower the q-value threshold). When comparing the output of TRIC to the naïve approach with a fixed q-value cutoff, we use these value computed during the FDR procedure also for the naïve data matrix (unless otherwise indicated).

Validation dataset

A random subset of 452 peptides were chosen from the *S. pyogenes* data described in Röst et al. [19] and 7,232 chromatograms were extracted across 16 LC-MS/MS runs. These 7,232 chromatograms were manually inspected using the Skyline software [20], the correct

peak (if present) was marked for each of the 16 runs and then exported. An in-house script compared the results of the manual annotation and the TRIC-based annotation where the peak was considered correct either if its apex differed less than 20 seconds from the manual annotation or of its apex was contained within the manual peak boundaries.

To study the correction of chromatographic distortion, a set of *S. pyogenes* runs acquired as part of a larger study was chosen for analysis and one run (`hroest_K131126_005`) was selected as the target and 7 additional runs from the same dataset (`hroest_K131126_050` to `hroest_K131126_056`) were selected for co-alignment (with about 45 other injections between the two acquisitions) to produce realistic alignment conditions. From the target run, half of all high confidence peakgroups (q-value < 0.001, 506 peakgroups) were removed. Then the full algorithm was run with the standard settings but `--alignment_score` changed to 0.001 to allow for more "anchor points". After running the algorithm on the "training" set, the result was compared to the retention times and intensities of the original 506 high confidence "test" peak groups that were set aside before.

Parameters used

Unless otherwise indicated, the `feature_alignment.py` (available at `https://pypi.python.org/pypi/msproteomicstools`) was run with the following settings: LocalMST for `--method` indicating the use of the MST, 0.0001 for the parameter `--alignment_score` only allowing highly confident peakgroups with q-value $< 0.01\,\%$ as "anchor points", lowess for `--realign_method`, 0.1 for `--max_fdr_quality` allowing for the transfer of confidence to peakgroups with a score cutoff of 0.1, a value of 0.01 for `--target_fdr` indicating $1\,\%$ FDR on the data matrix rows, True for `--mst:useRTCorrection` and 3.0 for `--mst:Stdev_multiplier` indicating adaptive retention time windows. The optional noise imputation algorithm implemented in `requantAlignedValues.py` was run with the following settings: singleClosestRun for `--method` and lowess for `--realign_method`.

pulsed-SILAC analysis

For the pulsed-SILAC analysis of the iPS cells, TRIC was run as described earlier. The `--max_fdr_quality` parameter was lowered to 0.05 (to only include higher quality peak

groups) and TRIC calculated a `m_score` cutoff of 0.002198 for the whole dataset. The data was filtered by `m_score` at the calculated cutoff (0.002198) to produce the data for the "naïve" approach (simple FDR filter) and filtered at 0.025 for the TRIC alignment data. To compare the difference in error between SILAC pairs, only points with an error less than 30 seconds were plotted in Figure 4 (148 high quality pairs and 165 aligned pairs were omitted as they fall outside the plotting window). At each time point, the amount of heavy (I_H) and light (I_L) precursor was extracted and used to calculated the relative isotopic abundance (RIA_t):

$$RIA_t = \frac{I_L}{I_L + I_H} \tag{1}$$

analogous to Pratt et al. [38] (with heavy and light switched due to our experimental design being reversed). The value of RIA_t is time dependent as unlabelled proteins are replaced with heavy-labelled proteins during the course of the experiment. This is due loss due to dilution of the cells as well as loss due to intracellular protein turnover, where the rate of loss can be modelled as an exponential decay process:

$$RIA_t = RIA_0 \cdot e^{-k_{loss} \cdot t} \tag{2}$$

where RIA_0 denotes the initial isotopic ratio and k_{loss} the rate of (hourly) loss of unlabelled protein. We assumed $RIA_0 = 1$ as no heavy isotope was present at $t = 0$, thus the value of RIA_t will decay exponentially from initially one to zero after infinite time. As discussed in Pratt et al. [38], these assumptions may reduce measurement error especially at the beginning of the experiment where isotopic ratios are more inaccurate due to the low absolute number of heavy precursor ions. Next, a linear model was fitted per peptide to the logarithmized data to obtain k_{loss} values for all individual peptides. The k_{loss} for each protein was computed as the median of all peptide-level rates. We excluded proteins quantified in a single timepoint only (241), proteins without a significant correlation ($p < 0.25$ of a linear model) between isotope ratio and time (90) and increasing isotope ratio over time (21). After filtering, 1075 proteins were used to compute turnover rates. In order to obtain the protein turnover rate $k_{turnover}$, we subtracted the dilution ratio D which we obtained by MTS assay from independent experiments on the same cell line at comparable confluence:

$$k_{turnover} = k_{loss} - D \tag{3}$$

Assuming the cells are in steady state and protein synthesis is equal to degradation, the computed $k_{turnover}$ is equal to the degradation rate: $k_{turnover} = k_{degradation}$. Please note for

the specific purpose of illustrating the TRIC alignment in this paper, we treated heavy and light channels separately, which could be further optimized (e.g., by combining heavy and light assays in the library generation step) for future SILAC experiments.

For gene ontology (GO) term enrichment, we used the Gene Ontology enrIchment analysis and visualization tool (GORILLA) at `http://cbl-gorilla.cs.technion.ac.il` as described in Eden et al. [39]. We selected the proteins with the highest and lowest turnover rates (10 % and 25 % quantiles) and used all identified proteins as background in GORILLA. We identified 20 significantly (q-value < 0.05) enriched GO terms. Most terms enriched for proteins with high turnover (fast degradation) were related to cell signaling, radiation and light response, cell-cell adhesion, extracellular matrix and locomotion (**Supplementary Table 6**).

Code availability

All source code for TRIC is available at `https://github.com/msproteomicstools/msproteomicstools` under the 3-clause BSD licence.

Data

All raw data and output generated by the tools described in this manuscript are deposited at the peptide atlas FTP site. It can be accessed at `ftp.peptideatlas.org` using the username PASS00788 and the password MP6824ws to log in or the full url `ftp://PASS00788:MP6824ws@ftp.peptideatlas.org/`.

Acknowledgements

This work was supported by ETH Zurich (grant ETH-30 11-2 to H.R. and R.A.), the Swiss National Science Foundation (SNSF grant P2EZP3_162268 to H.R.), ERC Proteomics v3.0 (grant 233226 to R.A.), the PhosphonetX project of SystemsX.ch, the ERC DISEASEAVATARS (grant 616441 to R.A. and G.T.), the Telethon Foundation (Grant n. GGP14265 to G.T.) and the Regione Lombardia (Grant Ricerca Indipendente 2012 to G.T.). Further funding was provided to R.A. by the Swiss National Science Foundation (SNSF).

We would like to thank SyBIT project of SystemsX.ch for support and maintenance of the lab-internal computing infrastructure. Specifically, L. Blum from SyBIT contributed substantially to make this software available to the whole lab and provided critical feedback.

Contributions

H.R. designed and wrote the code, performed the data analysis and produced the figures. Y.L., G.D. and M.Z performed the iPS experiment and acquired the mass spectrometric data. P.N. and G.R. contributed to the code and provided an initial prototype of the implementation. B.C. and L.G. acquired mass spectrometric data and gave critical input. G.T., L.M. and R.A. designed and supervised the study. All authors contributed to the manuscript.

Competing financial interests

The authors declare that they have no competing financial interests.

References

[1] Hudson, T. J. *et al.* International network of cancer genome projects. *Nature* **464**, 993–998 (2010).

[2] McLendon, R. *et al.* Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061–1068 (2008).

[3] Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).

[4] Haines, J. L. *et al.* Complement factor H variant increases the risk of age-related macular degeneration. *Science* **308**, 419–421 (2005).

[5] International Consortium for Blood Pressure Genome-Wide Association Studies and others. Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature* **478**, 103–109 (2011).

[6] Röst, H. L., Malmström, L. and Aebersold, R. Reproducible quantitative proteotype data matrices for systems biology. *Mol. Biol. Cell* **26**, 3926–3931 (2015).

[7] de Godoy, L. M. F. *et al.* Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature* **455**, 1251–1254 (2008).

[8] Hebert, A. S. *et al.* The one hour yeast proteome. *Mol. Cell. Proteomics* **13**, 339–347 (2014).

[9] Beck, M. *et al.* The quantitative proteome of a human cell line. *Mol. Syst. Biol.* **7**, 549 (2011).

[10] Nagaraj, N. *et al.* Deep proteome and transcriptome mapping of a human cancer cell line. *Mol. Syst. Biol.* **7**, 548 (2011).

[11] Desiere, F. *et al.* The peptideatlas project. *Nucleic Acids Res.* **34**, D655–D658 (2006).

[12] Omenn, G. S. *et al.* Metrics for the Human Proteome Project 2015: progress on the human proteome and guidelines for high-confidence protein identification. *J. Proteome Res.* **14**, 3452–3460 (2015).

[13] Picotti, P. *et al.* A complete mass-spectrometric map of the yeast proteome applied to quantitative trait analysis. *Nature* **494**, 266–270 (2013).

[14] Li, X.-j. *et al.* A blood-based proteomic classifier for the molecular characterization of pulmonary nodules. *Sci. Transl. Med.* **5**, 207ra142 (2013).

[15] Drabovich, A. P. *et al.* Differential diagnosis of azoospermia with proteomic biomarkers ECM1 and TEX101 quantified in seminal plasma. *Sci. Transl. Med.* **5**, 212ra160 (2013).

[16] Surinova, S. *et al.* Prediction of colorectal cancer diagnosis based on circulating plasma proteins. *EMBO Mol. Med.* e201404873 (2015).

[17] Surinova, S. *et al.* Non-invasive prognostic protein biomarker signatures associated with colorectal cancer. *EMBO Mol. Med.* **7**, 1153–1165 (2015).

[18] Gillet, L. C. *et al.* Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. *Mol. Cell. Proteomics* **11**, O111.016717 (2012).

[19] Röst, H. L. *et al.* OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. *Nat. Biotechnol.* **32**, 219–223 (2014).

[20] MacLean, B. *et al.* Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* **26**, 966–968 (2010).

[21] Martin, D. B. *et al.* MRMer, an interactive open source and cross-platform system for data extraction and visualization of multiple reaction monitoring experiments. *Mol. Cell. Proteomics* **7**, 2270–2278 (2008).

[22] Mead, J. A. *et al.* MRMaid, the Web-based Tool for Designing Multiple Reaction Monitoring (MRM) Transitions. *Mol. Cell. Proteomics* **8**, 696–705 (2009).

[23] Prakash, A. *et al.* Expediting the development of targeted SRM assays: using data from

shotgun proteomics to automate method development. *J. Proteome Res.* **8**, 2733–2739 (2009).

[24] Walsh, G. M. *et al.* Implementation of a data repository-driven approach for targeted proteomics experiments by multiple reaction monitoring. *J. Proteomics* **72**, 838–852 (2009).

[25] Sherwood, C. A. *et al.* MaRiMba: A Software Application for Spectral Library-Based MRM Transition List Assembly. *J. Proteome Res.* **8**, 4396–4405 (2009).

[26] Bertsch, A. *et al.* Optimal de novo Design of MRM Experiments for Rapid Assay Development in Targeted Proteomics. *J. Proteome Res.* **9**, 2696–2704 (2010).

[27] Reiter, L. *et al.* mProphet: automated data processing and statistical validation for large-scale SRM experiments. *Nature Methods* **8**, 430–435 (2011).

[28] Teleman, J. *et al.* Automated selected reaction monitoring software for accurate label-free protein quantification. *J. Proteome Res.* **11**, 3766–3773 (2012).

[29] Prakash, A. *et al.* Signal Maps for Mass Spectrometry-based Comparative Proteomics. *Mol. Cell. Proteomics* **5**, 423–432 (2006).

[30] Mueller, L. N. *et al.* SuperHirn - a novel tool for high resolution LC-MS-based peptide/protein profiling. *Proteomics* **7**, 3470–3480 (2007).

[31] Elias, J. E. and Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature Methods* **4**, 207–214 (2007).

[32] Reiter, L. *et al.* Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry. *Mol. Cell. Proteomics* **8**, 2405–2417 (2009).

[33] Teleman, J. *et al.* DIANA - algorithmic improvements for analysis of data-independent acquisition MS data. *Bioinformatics* (2014).

[34] Röst, H. L., Schmitt, U., Aebersold, R. and Malmström, L. Fast and Efficient XML Data Access for Next-Generation Mass Spectrometry. *PLOS One* (2015).

[35] Doherty, M. K., Hammond, D. E., Clague, M. J., Gaskell, S. J. and Beynon, R. J. Turnover of the human proteome: determination of protein intracellular stability by dynamic SILAC. *J. Proteome Res.* **8**, 104–112 (2009).

[36] Schwanhäusser, B. *et al.* Global quantification of mammalian gene expression control. *Nature* **473**, 337–342 (2011).

[37] Reinstein, E. and Ciechanover, A. Narrative review: protein degradation and human diseases: the ubiquitin connection. *Ann. Intern. Med.* **145**, 676–684 (2006).

[38] Pratt, J. M. *et al.* Dynamics of protein turnover, a missing dimension in proteomics. *Mol.*

Cell. Proteomics **1**, 579–591 (2002).

[39] Eden, E., Navon, R., Steinfeld, I., Lipson, D. and Yakhini, Z. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* **10**, 48 (2009).

[40] Liu, Y. *et al.* Quantitative variability of 342 plasma proteins in a human twin population. *Mol. Syst. Biol.* **11**, 786 (2015).

[41] Prince, J. T. and Marcotte, E. M. Chromatographic alignment of ESI-LC-MS proteomics data sets by ordered bijective interpolated warping. *Analytical Chemistry* **78**, 6140–6152 (2006).

[42] Kohlbacher, O. *et al.* TOPP–the OpenMS proteomics pipeline. *Bioinformatics* **23**, e191–197 (2007).

[43] Sturm, M. *et al.* OpenMS - an open-source software framework for mass spectrometry. *BMC Bioinformatics* **9**, 163 (2008).

[44] Cox, J. and Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367–1372 (2008).

[45] Tsou, C.-C. *et al.* DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. *Nature methods* **12**, 258–264 (2015).

[46] Adamo, A. *et al.* 7q11.23 dosage-dependent dysregulation in human pluripotent stem cells affects transcriptional programs in disease-relevant lineages. *Nature Genetics* **47**, 132–141 February (2015).

[47] Kim, S. C. *et al.* A clean, more efficient method for in-solution digestion of protein mixtures without detergent or urea. *J. Proteome Res.* **5**, 3446–3452 December (2006).

[48] Escher, C. *et al.* Using iRT, a normalized retention time for more targeted measurement of peptides. *Proteomics* **12**, 1111–1121 (2012).

[49] Liu, Y. *et al.* Glycoproteomic analysis of prostate cancer tissues by SWATH mass spectrometry discovers N-acylethanolamine acid amidase and protein tyrosine kinase 7 as signatures for tumor aggressiveness. *Mol. Cell. Proteomics* **13**, 1753–1768 July (2014).

[50] Collins, B. C. *et al.* Quantifying protein interaction dynamics by SWATH mass spectrometry: application to the 14-3-3 system. *Nature Methods* **10**, 1246–1253 (2013).

[51] Kessner, D., Chambers, M., Burke, R., Agus, D. and Mallick, P. ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics (Oxford, England)* **24**, 2534–2536 (2008).

[52] Craig, R. and Beavis, R. C. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **20**, 1466–1467 (2004).

[53] Geer, L. Y. *et al.* Open mass spectrometry search algorithm. *J. Proteome Res.* **3**, 958–964 (2004).

[54] Kunszt, P. *et al.* iPortal: the swiss grid proteomics portal: Requirements and new features based on experience and usability considerations. *Concurrency and Computation: Practice and Experience* **27**, 433–445 (2015).

[55] Keller, A., Eng, J., Zhang, N., Li, X. J. and Aebersold, R. A uniform proteomics MS/MS analysis platform utilizing open XML file formats. *Molecular Systems Biology* **1**, 17 (2005).

[56] Keller, A., Nesvizhskii, A. I., Kolker, E. and Aebersold, R. Empirical Statistical Model To Estimate the Accuracy of Peptide Identifications Made by MS/MS and Database Search. *Anal. Chem.* **74**, 5383–5392 (2002).

[57] Shteynberg, D. *et al.* iProphet: Multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates. *Mol. Cell. Proteomics* **10** (2011).

[58] Schubert, O. T. *et al.* Building high-quality assay libraries for targeted analysis of SWATH MS data. *Nat. Protoc.* **10**, 426–441 (2015).

[59] Lam, H. *et al.* Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics* **7**, 655–667 (2007).

Figures

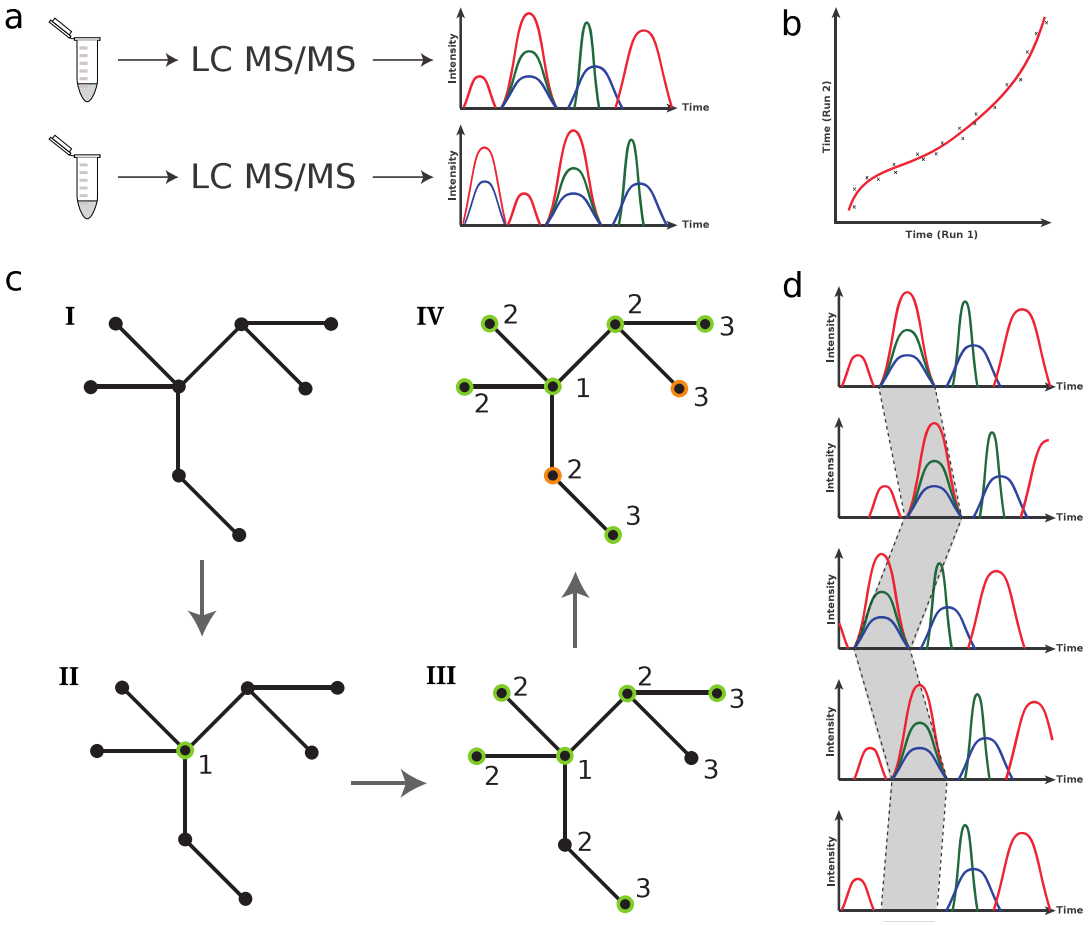

FIG. 1: **TRIC: Alignment algorithm for targeted proteomics data.** (**a**) In a targeted proteomics experiment, each run is typically analyzed individually, giving rise to multiple putative peak groups per run that may not be directly mappable due to chromatographic shifts. (**b**) The TRIC algorithm selects a set of high-confidence "anchor points" (peptides) for pairwise non-linear alignment and chromatographic distance estimation. (**c**) Based on the chromatographic distance, an optimal guidance tree (I) is computed (nodes are runs, edges are pairwise alignments). Next (II), the algorithm uses a starting point (1) to transfer identification confidence to nearby runs (iterations 2 and 3) using the guidance tree (III). In an optional last step (IV), runs without suitable peakgroups are re-visited to perform optional noise re-quantification (integration of all fragment ion signal at the aligned position is integrated; orange circles). (**d**) The confidence transfer step uses a starting peakgroup (top run) to select a narrow region in a neighboring run (gray region in second run) from which a peak gets selected. This procedure is repeated across all runs to identify the correct peak or establish peak boundaries in runs without any analyte signal (bottom run). In a real application, the alignment order may not be linear but follow the guidance tree.

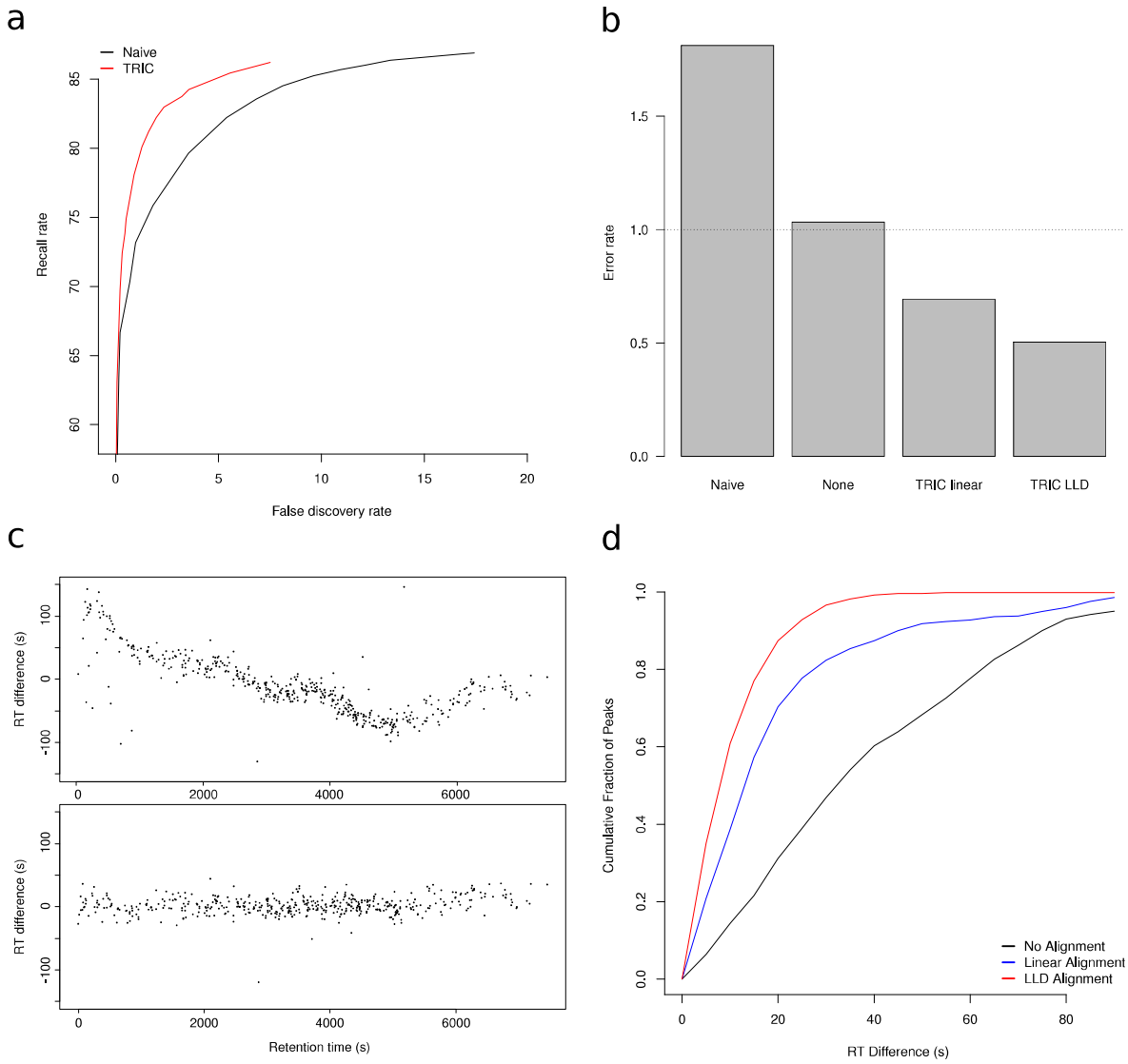

FIG. 2: **Identification and alignment accuracy of TRIC on manually annotated data.** We used a set of over 7,000 manually validated peakgroups to validate the TRIC algorithm. (**a**) FDR-Recall plot displaying recall rate *versus* the false discovery rate allows evaluation of the performance of TRIC compared to the naïve approach of using a fixed q-value cutoff applied to each run individually. As mis-classified peaks cannot be recovered even at high score cutoffs, a recall of 100 % cannot be reached. (**b**) Error rates at reported FDR cutoffs of 1 % for the naïve approach and TRIC without RT alignment (None), linear alignment (Linear) and non-linear k-nearest neighbor alignment (LLD). (**c**) The error of reported retention times are plotted without (top) and with (bottom) non-linear alignment on a sample run. (**d**) The cumulative fraction of peaks having less than a given error in retention time is plotted. TRIC with k-nearest neighbor smoothing (LLD) achieves high peak counts at low RT errors and outperforms linear or no alignment.

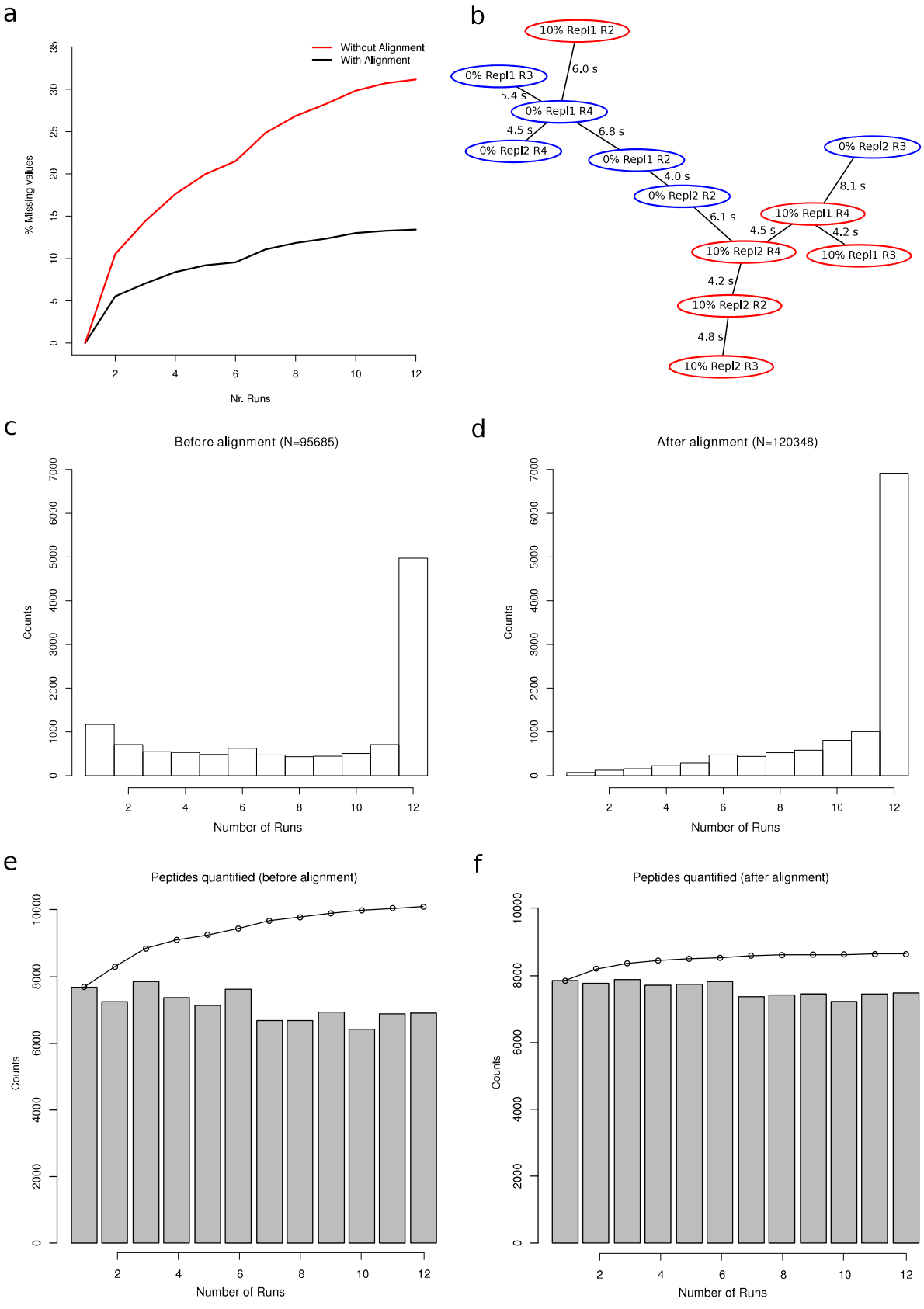
a
Without Alignment
With Alignment
% Missing values
Nr. Runs
b
10% Repl1 R2
6.0 s
0% Repl1 R3
5.4 s
0% Repl1 R4
4.5 s
6.8 s
0% Repl2 R4
0% Repl1 R2
4.0 s
0% Repl2 R2
6.1 s
0% Repl2 R3
8.1 s
10% Repl1 R4
4.5 s
4.2 s
10% Repl2 R4
10% Repl1 R3
4.2 s
10% Repl2 R2
4.8 s
10% Repl2 R3
c
Before alignment (N=95685)
Counts
Number of Runs
d
After alignment (N=120348)
Counts
Number of Runs
e
Peptides quantified (before alignment)
Counts
Number of Runs
f
Peptides quantified (after alignment)
Counts
Number of Runs

FIG. 3: **Analysis of a microbial dataset investigating *S. pyogenes* virulence.** A dataset of 12 runs of *S. pyogenes* exposed to human plasma was analyzed with TRIC. (**a**) The data matrix occupancy is higher after alignment with TRIC (fewer missing values are observed). (**b**) The computed guidance tree captures orthogonal information to injection order (root mean square deviation between runs is indicated for each edge). Control samples are in blue and plasma-exposed samples are in red (note that the tree is substantially different from injection order as samples were shot in three batches: R1, R2 and R3). (**c**) Number of precursors appearing in a specific number of runs before (left) and after (right) running TRIC; fully aligned precursors increased by 39 % while precursors found in only a single run decreased by 93.7 %. (**d**) The cumulative number of the number of peptides quantified using a fixed 0.01 q-value cutoff without alignment (left) and after applying TRIC and a minimal q-value cutoff of 0.0015 (right). While TRIC decreased the variance of the number of identifications across runs, the cumulative number of peptides also saturates more quickly indicating less accumulation of false positive identifications.

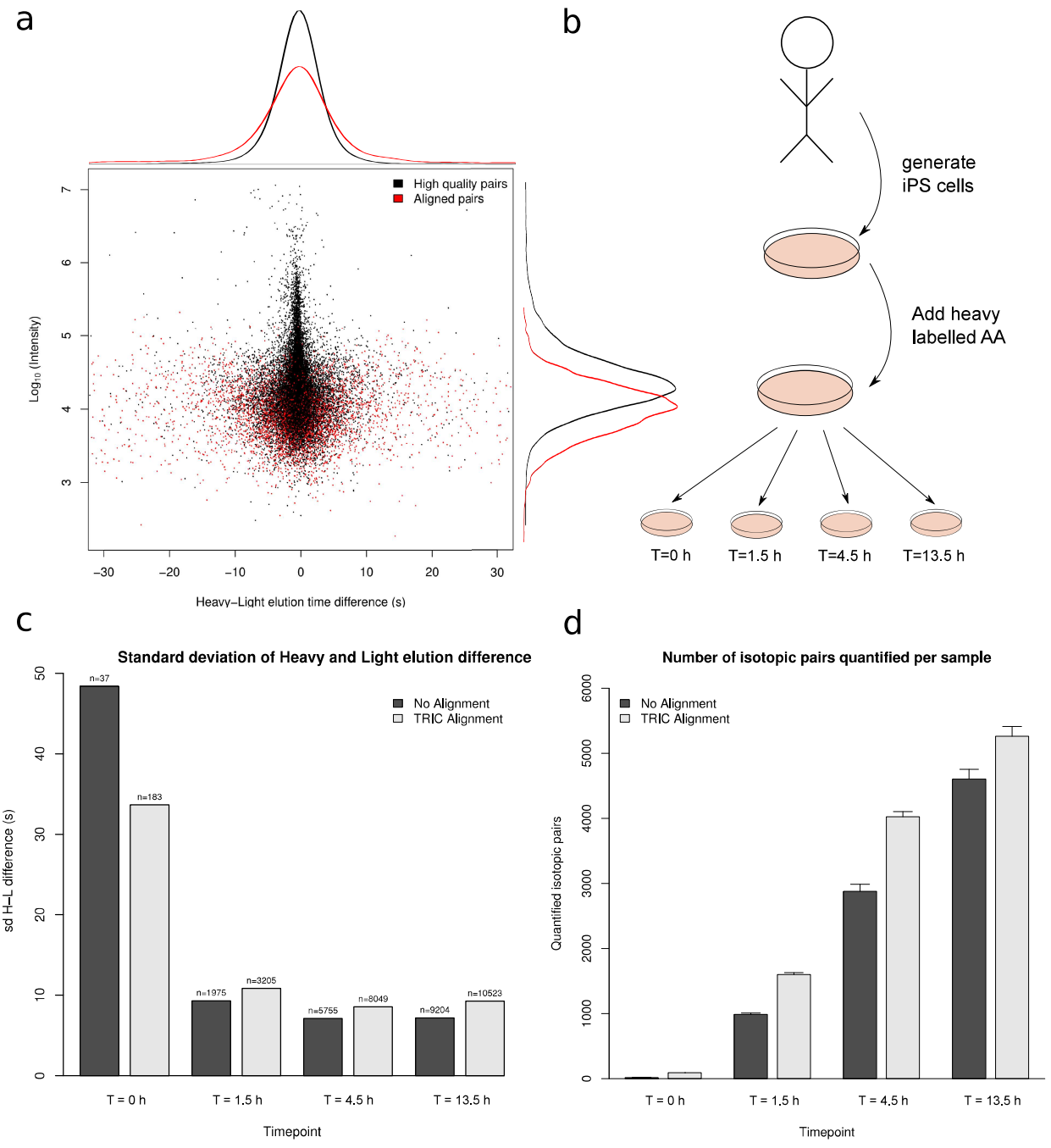
a
High quality pairs
Aligned pairs
Log10 (Intensity)
Heavy−Light elution time difference (s)
b
generate
iPS cells
Add heavy
labelled AA
T=0 h
T=1.5 h
T=4.5 h
T=13.5 h
c
Standard deviation of Heavy and Light elution difference
No Alignment
TRIC Alignment
sd H−L difference (s)
n=37
n=183
n=1975
n=3205
n=5755
n=8049
n=9204
n=10523
T = 0 h
T = 1.5 h
T = 4.5 h
T = 13.5 h
Timepoint
d
Number of isotopic pairs quantified per sample
No Alignment
TRIC Alignment
Quantified isotopic pairs
T = 0 h
T = 1.5 h
T = 4.5 h
T = 13.5 h
Timepoint

FIG. 4: **Pulsed-SILAC experiment performed on human iPS cells.** A human iPS cell line was exposed to a pulse of heavy amino acids and sampled at four time points in duplicates (see Panel **b**). (**a**) The RT difference between the light and heavy signal as a function of the intensity. Aligned values reported by TRIC (in red) have lower intensity and higher RT error (distribution on top only displays values below 10^4 in intensity) (**c**) Standard deviation of the RT difference between heavy and light pairs with and without TRIC alignment. For the analysis without alignment, a simple FDR cutoff was applied (naïve approach). Alignment increases the number of quantified SILAC pairs at the cost of slightly higher variance. Pairs from both replicates are aggregated. No heavy-light pairs are expected at $t = 0$ as heavy amino acids were added afterwards. (**d**) The number of isotopic SILAC pairs quantified per sample increases through the TRIC alignment, especially for the earlier time points with little heavy isotope signal. For each timepoint, average values across two replicates are shown with standard deviation.

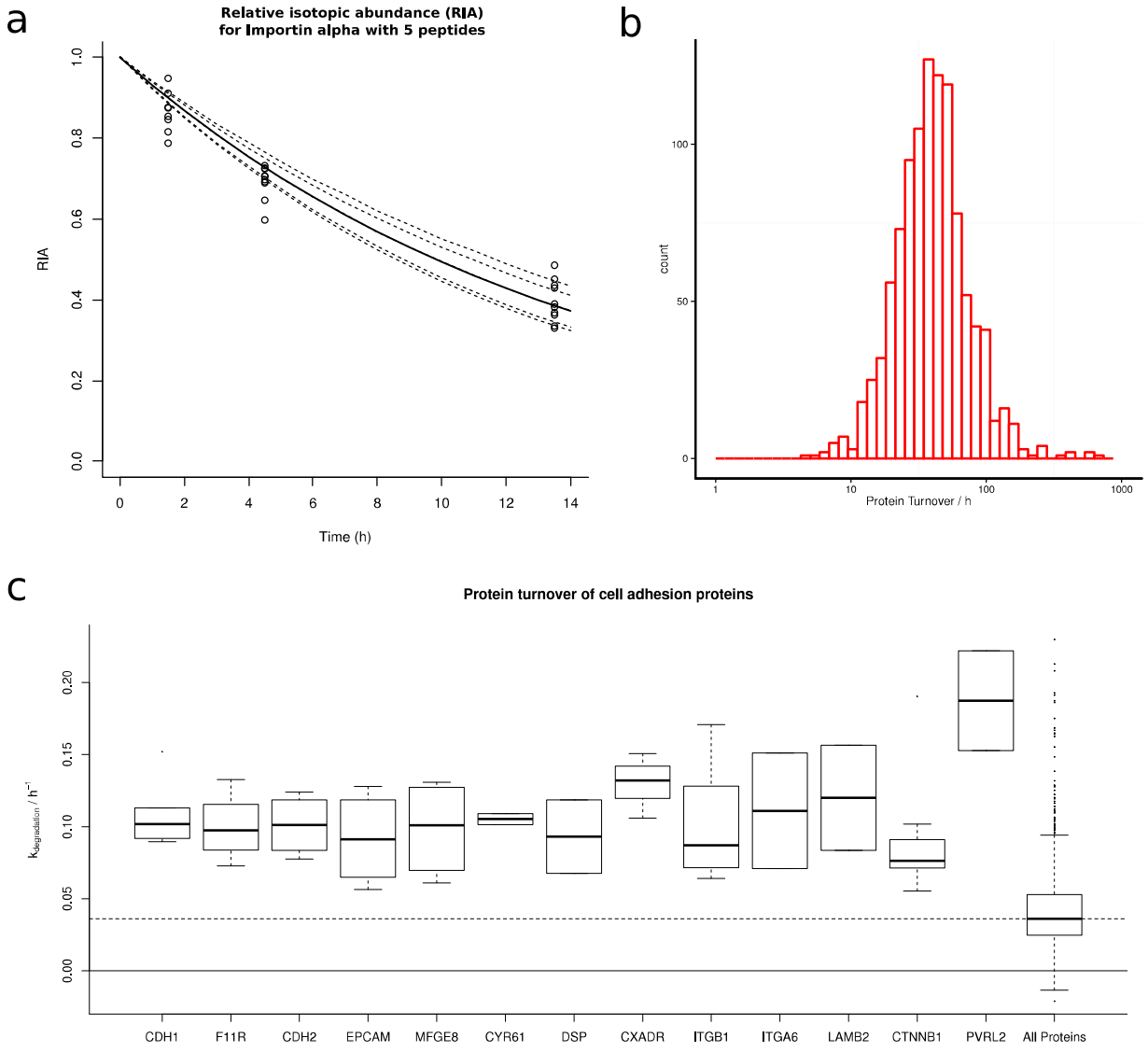

FIG. 5: **Protein turnover rates in human iPS cells.** Targeted proteomics analysis of protein turnover in human iPS cells. (**a**) Relative isotopic abundance (RIA) is plotted for an example protein, Importin Alpha, with 5 peptides (dashed lines). The median of all decay curves fitted through 1.0 at timepoint zero for all peptides estimates protein-level k_{loss}. (**b**) Global protein turnover rates are estimated after correction for protein dilution. (**c**) Proteins in GO category "cell adhesion proteins" show significantly higher turnover than expected ($p < 10^{-7}$). All peptides of the respective proteins exhibit substantially higher degradation rates than the base distribution (shown on the very right). Only proteins with two or more peptides are shown (box indicates first and third quartile with median shown in black; whiskers extend to the most extreme data point which is no more than 1.5 times the length of the box away from the box).